Deep Neural Networks for Parameter Estimation with Inverse Maps and for Subgrid-Scale Models on the Cerebras CS-2 AI-Cluster

Scientific Achievement
- Train large-scale neural networks on the Cerebras CS-2 AI accelerator to learn inverse maps that estimate parameters in physical and statistical models
- Design new convolutional neural networks enabling super-resolution for subgrid-scale variability of wind speeds in Earth’s atmosphere

Significance and Impact
Leverage new AI hardware to quickly train large-scale neural network models with tens of millions of trainable parameters. Training time on Cerebras CS-2 increases only mildly while the number of trainable parameters grows exponentially.

Research Details
- Performance evaluation of training on Cerebras CS-2 compared to a 4-core CPU system shows that large-scale models with tens of millions of parameters are trained extremely efficiently on the CS-2 platform
- Long setup phase of CS-2 is amortized by fast execution of training algo’s
- In order to transition to CS-2 system, operations in neural network models must be supported by the Cerebras framework

Table 4: Performance comparison between a reference CPU system and Cerebras CS-2. The number of units per layer is varied from 32 to 2048 while keeping the number of layers fixed to eight; the trainable NN parameters increase exponentially. The runtimes are given for 1,000 epochs or processing of a total of 1,000,000 training samples.

<table>
<thead>
<tr>
<th>Units per layer</th>
<th>CPU train</th>
<th>CS-2 setup</th>
<th>CS-2 train</th>
<th>Total [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>4,9</td>
<td>4.6</td>
<td>4.9</td>
<td>31M</td>
</tr>
<tr>
<td>128</td>
<td>4,9</td>
<td>4.6</td>
<td>4.9</td>
<td>31M</td>
</tr>
<tr>
<td>512</td>
<td>4,9</td>
<td>4.6</td>
<td>4.9</td>
<td>31M</td>
</tr>
<tr>
<td>2048</td>
<td>4,9</td>
<td>4.6</td>
<td>4.9</td>
<td>31M</td>
</tr>
</tbody>
</table>

This performance study has the purpose to stress the data throughput of the CS-2 system and compare it for training, which increases only mildly while NN models increase exponentially in the number of trainable parameters.

Figure 2: Comparing runtimes of a reference CPU system and Cerebras CS-2. CS-2 has a very low runtime for subgrid-scale variability for estimating parameters in physical and statistical models.

Figure 4: Super-Resolution for Subgrid-Scale Variability. This performance study has the purpose to stress the data throughput of the CS-2 system and compare it for training.

Results for Subgrid-Scale Variability
This performance study has the purpose to stress the data throughput of the CS-2 system and compare it for training.

Figure 3: Performance evaluation of training on Cerebras CS-2 compared to a reference CPU system. The figure has two rows, each corresponding to one sample of the testing data set. The left column of plots shows the coarse resolution images of wind speeds that are input into the CNN and the subsequent column shows the NN outputs in fine resolutions. The true wind speeds are shown in the third column and the rightmost column depicts the errors between true and predicted. The results were obtained on the reference CPU system. The figure has two columns, one for accuracy and the other for timing.

- High accuracy with neural network model with 31M trainable parameters
- Low accuracy with neural network model with 39K trainable parameters